Exam 2 (100 pts)

Name:_____**KEY**_____

1. Which of following is a reasonable estimate of the size of the E. coli genome?
    a. ~4 bp
    b. ~4 Gbp (4,000,000,000 base pairs)
    c. ~4 Tbp (4,000,000,000,000 base pairs)
    d. None of the above

2. Which of the following is a reasonable estimate of size the grizzly bear's genome?
    a. ~3 Mbp (3,000,000 base pairs)
    b. ~3 Gbp (3,000,000,000 base pairs)
    c. ~30 Gbp (30,000,000,000 base pairs)
    d. ~30 Tbp (30,000,000,000,000 base pairs)

3. What feature is required for assembling short sequences into longer contigs?
    a. The sequences must overlap
    b. The sequences must be the same length
    c. The sequences must start with the same 5' nucleotide
    d. All of the above

4. Sequencing the human genome using high-throughput sequencing is much faster than sequencing the human genome using conventional sequencing methods:
    a. True
    b. False

5. Sequencing the human genome using high-throughput sequencing is much cheaper than sequencing the human genome using conventional sequencing methods:
    a. True
    b. False

6. Conventional sequencing approaches that are not considered high-throughput are no longer used in research:
    a. True
    b. False

7. Which of the following high-throughput sequencing methods is most common?
    a. Ion Torrent
    b. Illumina
    c. PacBio
    d. Nanopore

8. What is the difference between single end and paired end sequencing?
    a. Single end sequencing doesn't involve bridge amplification and cluster generation
    b. Single end sequencing is done on a flow cell and paired end sequencing is done on a plate
    c. Single end sequencing involves sequencing from only one end of a fragment of DNA
    d. All of the above

9. What features could influence your decision on which sequencing technology to use:
    a. Throughput
    b. Read length
    c. Error rate
    d. All of the above

10. Which of the following is a typical size range for short read sequencing technologies?
    a. <10 nt
    b. 10-50 nt
    c. 50-500 nt
    d. 1,000-10,000 nt
    e. 10,000 -100,000 nt

11. Which of the following Phred scores would be used as a cutoff to discard reads from a Next-gen sequencing run that have a probability of >1 error per 1,000 bases?
    a. 10
    b. 20
    c. 30
    d. 40
    e. 50

12. You're interested in sequencing a mutant strain of Arabidopsis to identify the causal mutation for a particular phenotype. You determine that you need 20X coverage of the genome and the genome is 100 million base pairs. If your read length is 100 bases, how many reads would you need to sequence the genome at 20X coverage?

    Summary
    Coverage: 20X
    Read length: 100 bases
    Genome: 100,000,000 basepairs

    reads = 20 x 100,000,00 / 100 => 20,000,000

13. Suppose you are guaranteed at least 300 million reads from each lane of a flowcell and you need 20 million reads for each of your samples. How many samples can you multiplex into one lane?
    a. 5
    b. 10
    c. 15
    d. 20
    e. 25

14. What common file format for high-throughput sequencing data is shown below:

    fastq

```
@D64TDFP1:248:C50DMACXX:5:1101:1241:2095 1:N:0:ATCACG
CACCGCCCGTCGCTATCCGGGACTGGAATTCTCGGGTGCCAAGGAACTCCA
+
CCCFFFFFHHHHHJIJGHJJJJIJJJJJGGGFFFFEABDHHHFHFF@@DD>
@D64TDFP1:248:C50DMACXX:5:1101:1371:2154 1:N:0:ATCACG
TCAATATTTGCATAGGGTATCTGGAATTCTCGGGTGCCAAGGAACTCCAGT
+
CCCFFFFFHHHHHJJJJGFHIJJJJJJJJJJJJJFHHIIJJHGHJFGHJJI
@D64TDFP1:248:C50DMACXX:5:1101:1461:2205 1:N:0:ATCACG
GAAAGACGTCTTCCTAGATTATGGAATTCTCGGGTGCCAAGGAACTCCAGT
+
CCCFFFFFHHHHHJJJJJJJJJJJIJJJJJJJJJHIJJJJJGIIJFGIJJJ
```

15. How many reads are represented in a file like the one shown in question 15 that contains 400 million lines of text?
    100,000,000

16. What information is contained in each of the following lines of the file snippet in question 15:

    Line 1: sequence ID

    Line 2: sequence

    Line 4: quality scores

17. What software is commonly used to assemble bacterial genomes?

    Velvet

18. What software is commonly used to map reads to a reference genome?

    Bowtie2

19. De Bruijn graphs are commonly used in genome assembly software:
    a. True
    b. False

20. What GUI is commonly used to assess library quality?

    FastQC

21. What information is contained in a SAM file that is not contained in a raw high-throughput sequencing data file?

    Position at which each sequence maps to a reference.

22. What is an N50 value in reference to genome assembly?
    a. The percentage of the genome covered by contigs.
    b. The contig length at which 50% of the assembled genome is in contigs of N50 or greater.
    c. The average contig length.
    d. The maximum contig length.

23. What file extension is associated with a GNU zipped (gzip compressed) file?

    .gz

24. Which of the following is a common genome editing technology used to introduce specific modifications or random mutations into a genome?
    a. Krispy Kreme
    b. CRISPER
    c. CRISPR
    d. KFC

25. What two components are required for CRISPR-based genome editing assuming random mutations are acceptable?

    gRNA and Cas9

26. In addition to the two components listed above, what is a third component required for engineering specific mutations and for introducing transgenes such as GFP into a genome using CRISPR?

    Repair template for homologous recombination

27. In which of the following organisms are you most likely to find endogenous CRISPR-based immune systems?
    a. Archaea and bacteria
    b. Mammals
    c. Birds
    d. Worms
    e. Sea anemones

28. The majority of physical traits in humans can be ultimately linked to a single genetic locus:
    a. True
    b. False

29. The majority of diseases can be linked to a single genetic locus:
    c. True
    d. False

30. Match the following terms with their definitions:
    a. Functional genomics
    b. Comparative genomics
    c. Genome-wide association studies

_a_ Application of high-throughput and genomics approaches to describe gene function
_c_ An observational study of genetic variation to identify variants associated with a particular trait
_b_ Comparison of genomics features across different species

31. What are regular expressions?

A sequence of characters that define a pattern

32. What is an example of a regular expression and what does it stand for?

\n = new line

33. How would you extract all lines of a fasta file called genes.fa that contain a start codon (ATG) and redirect the output to a new file called sequences.txt from the command line? Write your code next to the prompt.

$ grep ATG <genes.fa >sequences.txt